*Econometric Theory*, **16**, 2000, 551–575. Printed in the United States of America.

# SEMIPARAMETRIC ESTIMATION OF MULTIPLE EQUATION MODELS

GABRIEL A. PICONE
*University of South Florida*

J.S. BUTLER
*Vanderbilt University*

This paper proposes a semiparametric estimator for multiple equations multiple index (MEMI) models. Examples of MEMI models include several sample selection models and the multinomial choice model. The proposed estimator minimizes the average distance between the dependent variable unconditional and conditional on an index. The estimator is $\sqrt{N}$-consistent and asymptotically normally distributed. The paper also provides a Monte Carlo experiment to evaluate the finite-sample performance of the estimator.

## 1. INTRODUCTION

In recent years semiparametric estimators for limited dependent and qualitative variable models have attracted an increasing amount of interest because they do not require specification of the distribution function of the error terms but still have desirable statistical properties. When the distribution function of the error terms is unknown many limited dependent and qualitative variable models belong to the class of semiparametric index models. In these models, the explanatory variables influence the dependent variable only as an unknown function of a known index or indices.

Whereas several semiparametric estimators have been proposed to estimate single equation models with index restrictions, including Ichimura (1993) and Ichimura and Lee (1991), many econometric models involve the estimation of multiple equations with multiple indices. Examples include sample selection models with multiple equations and multiple selection terms and the multinomial choice model. Although these models could be estimated equation by equation using a multiple index single equation estimator, an estimator that accounts for the cross-equations restrictions and the variance-covariance matrix that these models imply will be more efficient.[1] This paper presents such an estimator. We extend the semiparametric least squares technique originally pro-

We thank James Foster, Brad Kamp, Jeff Racine, and Mark Wilson for helpful comments and also Donald Andrews for suggestions and encouragement on an early stage of the paper. We also thank a co-editor, Joel Horowitz, and an anonymous referee for substantial comments that greatly improved the quality of the paper. We are solely responsible for any errors. Address correspondence to: Gabriel Picone, Department of Economics, BSN3403, University of South Florida, 4202 East Fowler Ave., Tampa, FL 33620, USA; e-mail: gpicone@coba.usf.edu.

 0266-4666/00 $9.50

posed by Ichimura (1993) for the single index model and used by Ichimura and Lee (1991) for the multiple index single equation models to encompass models of multiple equations with multiple indices.

This paper develops a semiparametric weighted least squares (WLS) estimator for multiple equations with multiple index models. We prove that the semiparametric-WLS estimator is $\sqrt{N}$-consistent and asymptotically normally distributed. Whereas previous work on semiparametric index estimators assumed the regressors and the dependent variables were independent and identically distributed (i.i.d.), our asymptotic properties are shown to hold under the assumption that they are nonidentically distributed and $m$-dependent (independent beyond a lag of $m$ periods). Thus, the applicability of these results is extended to cases where the data exhibit some temporal dependence. We also derive the variance-covariance matrix and discuss the identification of the semiparametric WLS. Finally, in a Monte Carlo analysis we illustrate the finite-sample behavior of the estimator and compare its performance to other alternative parametric and semiparametric estimators when the distribution function is correctly specified and also when it is misspecified.

The remainder of this paper is organized as follows. Section 2 presents the multiple equations with multiple index model and useful examples of this model. Section 3 introduces the semiparametric-WLS estimator. Section 4 demonstrates asymptotic properties of the estimator and derives the variance-covariance matrix. Section 5 discusses asymptotic efficiency and shows that the parameters of each index are identified up to a multiplicative constant when the index is linear. Section 6 examines the estimator's finite-sample properties via Monte Carlo simulations, and Section 7 concludes the paper.

## 2. MULTIPLE EQUATIONS WITH MULTIPLE INDEX MODELS

We define a multiple equations with multiple index model (MEMI) as

$$\begin{aligned} Y_{1i} &= h_{10}(\theta_0, X_i) + F_1(h_1(\theta_0, X_i)) + \epsilon_{1i}, \\ &\vdots \\ Y_{mi} &= h_{m0}(\theta_0, X_i) + F_m(h_m(\theta_0, X_i)) + \epsilon_{mi}, \end{aligned} \tag{1}$$

where $Y_{ji}(j = 1, \ldots, m)$ is the dependent variable, $X_i(i = 1, \ldots, N)$ is a vector of exogenous random variables, $\theta_0$ is an unknown parameter vector, and $\epsilon_{ji}$ is a random disturbance with an unknown distribution function. The functions $h_{j0}: \mathbf{R}^k \times \Theta \to \mathbf{R}$ and $h_j: \mathbf{R}^k \times \Theta \to \mathbf{R}^{l_j}$ are known up to the parameter $\theta$ and are called indices. The function $F_j: \mathbf{R}^{l_j} \to \mathbf{R}$ is not known. We assume that $E(\epsilon_{ji}|X_i) = 0$ and $\mathrm{Var}(\epsilon_i|X_i) = \Omega_i(X_i, \theta_0)$. The covariance matrix $\Omega_i$ depends on the model being analyzed.

This definition allows for the dimension of the indices to be different in each equation. It also accounts for cross-equations parameter restrictions by allowing for the indices to be known functions of $\theta$. The functions $h_{j0}$ are included to make the model more general, and in many situations they will be equal to

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

zero. A large class of econometric models belongs to the class of MEMI models, including several sample selection models and the multinomial choice model. To illustrate the class of MEMI models, consider the following sample selection and truncated Tobit model:

$$\begin{aligned} Y_{1i}^* &= X_{1i}'\beta_1 + u_{1i}, \\ Y_{2i}^* &= X_{2i}'\beta_2 + u_{2i}, \\ Y_{3i}^* &= X_{3i}'\beta_3 + u_{3i}. \end{aligned} \tag{2}$$

We do not observe $Y_{1i}^*$, $Y_{2i}^*$, or $Y_{3i}^*$; instead we observe $Y_{1i} = Y_{1i}^*$ and $Y_{2i} = Y_{2i}^*$ if and only if $Y_{2i}^* > 0$ and $Y_{3i}^* > 0$. Thus, our sample is truncated. It can be shown that the observed random vectors $Y_1$ and $Y_2$ can be written as

$$\begin{aligned} Y_1 &= X_1'\beta_1 + F_1(X_2'\beta_2, X_3'\beta_3) + \epsilon_1 \\ Y_2 &= F_2(X_2'\beta_2, X_3'\beta_3) + \epsilon_2 \end{aligned} \tag{3}$$

where $F_1 : \mathbf{R}^2 \to \mathbf{R}$ and $F_2 : \mathbf{R}^2 \to \mathbf{R}$ are known functions only if the distribution function of the error terms is known, otherwise they are unknown functions of a multiple index. The error terms are defined as $\epsilon_1 = Y_1 - X_1'\beta_1 - F_1(\cdot)$ and $\epsilon_2 = Y_2 - F_2(\cdot)$. Thus, this model can be regarded as a MEMI model with two equations, where $h_{10}(\theta_0, X_i) = X_1'\beta_1$, $h_{20}(\theta_0, X_i) = 0$ and $h_j(\theta_0, X_i) = (X_2'\beta_2, X_3'\beta_3)$ for $j = 1,2$. The variance-covariance matrix will be equal to

$$\Omega_i = \begin{bmatrix} V(u_{1i}|u_{2i} > -X_2'\beta_2, u_{3i} > -X_3'\beta_3) & \operatorname{Cov}(u_{1i}u_{2i}|u_{2i} > -X_2'\beta_2, u_{3i} > -X_3'\beta_3) \\ \operatorname{Cov}(u_{1i}u_{2i}|u_{2i} > -X_2'\beta_2, u_{3i} > -X_3'\beta_3) & V(u_{2i}|u_{2i} > -X_2'\beta_2, u_{3i} > -X_3'\beta_3) \end{bmatrix}.$$

This model can be extended to allow for extra dependent variables and selection terms. For example, if $Y_{1i}^*$ is a $d_1$-vector, $Y_{2i}^*$ is a $d_2$-vector, and $Y_{3i}^*$ is a $d_3$-vector, then we will have $d_1 + d_2$ dependent variables and $d_2 + d_3$ selection terms. The model is also applicable when the sample is censored, rather than truncated.

Utility maximizing models with discrete and continuous choices such as the one presented in Dubin and McFadden (1984) also belong to the class of MEMI models. For example, assume that a consumer faces $M$ mutually exclusive choices. Let $I_{mi}^* = Z_{mi}\gamma + \eta_{mi}$ be the indirect utility for alternative $m$, $m = 1,\ldots,M$ and consumer $i$, $i = 1,\ldots,N$. We do not observe $I_{mi}^*$ but an indicator variable $I_{mi} = 1$ if $I_{mi}^* = \max(I_{1i}^*,\ldots,I_{Mi}^*)$ and 0 otherwise. Conditional on individual $i$ selecting the $m$th category we also observe $Y_{mi} = X_{mi}\beta_m + u_{mi}$ where $Y_{mi} \in \mathbf{R}^d$ is a vector of continuous dependent variables. In Dubin and McFadden's example a consumer chooses an appliance portfolio, $I_{mi}$, and given this choice we observe the consumption of electricity and an alternative energy source, $Y_{mi}$ with $d = 2$. Assume that $E(u_m|X_m, Z_m) = E(\eta_m|X_m, Z_m) = 0$ but the distribution function of the error terms is unknown. As discussed in Maddala (1983) this model can be written as

$$Y_{mi} = X_{mi}\beta_m + \varphi_m((Z_{mi} - Z_{1i})\gamma,\ldots,(Z_{mi} - Z_{Mi})\gamma) + \epsilon_{mi}, \tag{4}$$

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

and each of the resulting $M$ system of $d$ equations can be seen as a MEMI model. Alternatively, we could write all $M$ equations together as a system of seemingly unrelated nonlinear equations and it will also be a MEMI model with $d \times M$ equations and $M - 1$ indices. The following section provides a semiparametric estimator for the class of MEMI models.

## 3. A SEMIPARAMETRIC-WLS ESTIMATOR

The model given in (1) belongs to the class of seemingly unrelated nonlinear equation models. An analyst who knows the function $F = (F_1, \ldots, F_m)$ and the covariance matrix $\Omega_i$ could estimate $\theta_0$ by minimizing the following weighted sum of squared residuals (WLS):

$$S_N(\theta) = \sum_{i=1}^{N} (Y_i - h_0(\theta, X_i) - F(h(\theta, X_i)))' \Omega_i^{-1} (Y_i - h_0(\theta, X_i) - F(h(\theta, X_i))).$$

Because the functions $F_1(), \ldots, F_m()$ and $\Omega_i^{-1}$ are unknown the proposed WLS cannot be used. However, we can still estimate model (1) by combining nonparametric estimates for the unknown functions $F_1(), \ldots, F_m()$ and $\Omega_i^{-1}$ with the WLS just described.

The index $h_j(\theta, X_i)$ is an $l_j \times 1$ random vector with distribution function $f_j(h_j(\theta, X_i); \theta)$ dependent on the distribution function of $X_i$ and the value of $\theta$. Then, given $\theta$, the regression function of $(Y_{ji} - h_{j0}(\theta, X_i))$ on $h_j(\theta, X_i)$ evaluated at any point $v_j$ is equal to $r_j(v_j; \theta) = E((Y_{ji} - h_{j0}(\theta, X_i)) | h_j(\theta, X_i) = v_j)$ and is called the index regression function because as the value of $\theta$ changes the distribution function of $h_j(\theta, X_i)$ will change and the value of $r_j(v_j; \theta)$ will also change. When $\theta = \theta_0$ the index regression function $r(h_j(\theta_0, X_i); \theta_0)$ is equal to the unknown function $F_j(h_j(\theta_0, X_i))$. In general, for any other $\theta \neq \theta_0$, $F_j(h_j(\theta, X_i)) \neq r_j(h_j(\theta, X_i); \theta)$ for $j = 1, \ldots, m$.

We estimate the index regression function $r_j(h_j(\theta, X_i); \theta)$ using a nonparametric estimator for the regression function. There are several nonparametric estimators for $r_j(h_j(\theta, X_i); \theta)$ that could be used. In this paper we use the Nadaraya–Watson (N–W) kernel estimator for the index regression function. The N–W kernel estimator for $l_j$ multivariate regression functions evaluated at $v_{ji}$ is equal to

$$\hat{r}_j(v_{ji}; \theta) = \frac{\sum_{s \neq i}^{N} (Y_{js} - h_{j0}(\theta, X_s)) \hat{K}_j \left( \dfrac{v_{ji} - h_j(\theta, X_s)}{a_{jN}} \right)}{\sum_{s \neq i}^{N} \hat{K}_j \left( \dfrac{v_{ji} - h_j(\theta, X_s)}{a_{jN}} \right)}$$

with $\hat{K}_j(x) = \det(\hat{D}_j)^{-1/2} K_j(\hat{D}_j^{-1/2} x)$,

where $K_j()$ is the kernel function that is a nonrandom real function on $\mathbf{R}^{l_j}$, the $l_j \times l_j$ matrix $\hat{D}_j$ is a data-dependent scale matrix, and $a_{jN}$ is a smoothing pa-

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

rameter, also called the bandwidth. The N–W kernel estimator of the index regression function is widely used in the semiparametric literature and has the advantage that its asymptotic properties are well known and that, if the kernel function is differentiable, then the N–W kernel estimator is also differentiable.

This paper proposes the following weighted least squares method to estimate a MEMI model:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} I(X_i \in X_N)(\tilde{Y}_i - \hat{r}(h(\theta, X_i);\theta))'\hat{\Omega}_i^{-1}(\tilde{Y}_i - \hat{r}(h(\theta, X_i);\theta)), \quad \textbf{(5)}$$

where

$$\tilde{Y}_i = \begin{bmatrix} Y_{1i} - h_{01}(\theta, X_i) \\ \vdots \\ Y_{mi} - h_{0m}(\theta, X_i) \end{bmatrix}, \qquad \hat{r}(h(\theta, X_i);\theta) = \begin{bmatrix} \hat{r}_1(h_1(\theta, X_i);\theta) \\ \vdots \\ \hat{r}_m(h_m(\theta, X_i);\theta) \end{bmatrix},$$

$$h(\theta, X_i) = \begin{bmatrix} h_1(\theta, X_i) \\ \vdots \\ h_m(\theta, X_i) \end{bmatrix},$$

and $\hat{\theta}$ is called the semiparametric-WLS estimator. The term $\hat{\Omega}_i^{-1}$ is a consistent estimator of $\Omega_i^{-1}$ which is the optimal weighting matrix in the sense of providing the smallest asymptotic variance. The expression $\hat{\Omega}_i^{-1}$ depends on the model being studied and in general will be a nonparametric function of the index $h(\theta_0, X_i)$ as in the models previously discussed. Because $\theta_0$ is unknown, one should first obtain a consistent estimator of $\theta_0$ (call it $\tilde{\theta}$) by substituting the identity matrix or any symmetric positive semidefinite matrix for $\hat{\Omega}_i^{-1}$ in (5) and then use $\tilde{\theta}$ to obtain a nonparametric estimate of $\Omega_i^{-1}$. The trimming term $I(X_i \in X_N)$, where $I(\cdot)$ is the indicator function and $X_N$ is a nonstochastic and bounded subset of $\mathbf{R}^k$, is introduced to obtain uniform consistency of the index regression function and its derivatives with a desirable rate of convergence.

## 4. ASYMPTOTIC PROPERTIES

In this section we show the asymptotic properties of the semiparametric-WLS estimator. To simplify the proofs, we follow a two step procedure. In the first step, we show that the value of $\theta$ (call it $\tilde{\theta}$) that minimizes

$$\hat{Q}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} I(X_i \in X_N)(\tilde{Y}_i - \hat{r}(h(\theta, X_i);\theta))'A_i(\tilde{Y}_i - \hat{r}(h(\theta, X_i);\theta)) \quad \textbf{(6)}$$

has desirable asymptotic properties, where $A_i$ can be any symmetric, nonstochastic, positive semidefinite matrix. The proof of our asymptotic results can be modified to allow for a stochastic matrix, $\hat{A}_i$, if $\hat{A}_i \xrightarrow{p} A_i$ and it is a bounded random matrix. In the second step, we replace $A_i$ with $\hat{\Omega}_i^{-1}$.

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

We use the following notation. Let $\mu$ denote a $k$-vector of nonnegative integers.

(a) $|\mu| = \sum_1^k \mu_j$,
(b) for any function $g(x)$ on $\mathbf{R}^k$, $D^\mu g(x) = \partial^{|\mu|}/(\partial x_1^{\mu_1},\ldots,\partial x_k^{\mu_k})g(x)$, and
(c) $x^\mu = \prod_1^k x_j^{\mu_j}$.

The following assumptions are sufficient to ensure the asymptotic normality and $\sqrt{N}$-consistency of the estimator $\tilde{\theta}$ obtained by minimizing (6).

Assumption 1. $\{(Y_i, X_i) : i = 1,\ldots,N\}$ is a sequence of $m$-dependent random vectors, where $Y_i \in \mathbf{R}^m$ and $X_i \in \mathbf{R}^k$.

Assumption 2. $\Theta \subset \mathbf{R}^{k^*}$ is compact, and $\theta_0$ is an interior point of $\Theta$.

Assumption 3. $h_j(\theta_0, X_i)$ is twice continuously differentiable on $\theta$, and $E\|Y - h_0(\theta_0, X_i)\|^r < \infty$, $E\|(\partial/\partial\theta)h(\theta_0, X_i)\|^r < \infty$, $E\|(\partial/\partial\theta)h_0(\theta_0, X_i)\|^r < \infty$, and $E\|\epsilon_i\|^r < \infty$ with $r \geq 8$.

Assumption 4.

(a) The support of $h_j(\theta, X_i)$ for $j = 1,\ldots,m$ is $\mathbf{R}^{l_j}$, and the distribution of $h_j(\theta, X_i)$ is absolutely continuous with respect to Lebesgue measure with density $f_{ji}(v_j;\theta)$ $\forall\theta, j, i$.
(b) $f_{ji}(v_j, \theta)$ is continuously differentiable in $v_j$ to order $w_j > l_j + 2$ on $\mathbf{R}^{l_j}$ $\forall\theta$ and $\sup_{\mathbf{R}^{l_j}\times\Theta}|D^\mu f_{ji}(v_j;\theta)| < \infty$ $\forall\mu$ with $|\mu| \leq w_j$.

Assumption 5.

(a) $r_j(v_j;\theta) = E(Y_{ji} - h_{j0}(\theta, X_i)|h_j(\theta, X_i) = v_j)$ does not depend on $i$, $\forall\theta$.
(b) $r_j(v_j;\theta)f_{ji}(v_j;\theta)$ is continuously differentiable in $v_j$ to order $w_j > l_j + 2$ on $\mathbf{R}^{l_j}$ $\forall\theta$ and $\sup_{\mathbf{R}^{l_j}\times\Theta}|D^\mu[r_j(v_j;\theta)f_{ji}(v_j;\theta)]| < \infty$ $\forall\mu$ with $|\mu| \leq w_j$.
(c) $r_j(v_j;\theta)$ is twice continuously differentiable in $\theta$.

Assumption 6.

(a) The trimming set, $X_N$, is a nonstochastic and bounded subset of $\mathbf{R}^k$. The $\lim_{N\to\infty} X_N = X$, where $X$ is a compact set.
(b) The set $X$ is chosen such that $\inf_{X\times\Theta} f_{ji}(h_j(\theta, X_i);\theta) > 0$ for $\forall j, i$.

Assumption 7.

(a) Each kernel function $K_j(x)$ used in $\hat{r}_j(v_j;\theta)$ for $j = 1,\ldots,m$ satisfies the following conditions: $\int K_j(x)\,dx = 1$, $\int x^\mu K_j(x)\,dx = 0$, $\forall 1 \leq |\mu| \leq l_j + 1$, $\int |x^\mu K_j(x)|\,dx < \infty$ $\forall\mu$ with $|\mu| = l_j + 2$, $D^\mu K_j(x) \to 0$ as $\|x\| \to \infty$ $\forall\mu$ with $|\mu| = \max(l_j/2, 3)$ and $\sup_x |D^{\mu+e_i}K_j(x)|(\|x\| \vee 1) < \infty$ $\forall\mu$, with $|\mu| = \max(l_j/2, 3)$ $\forall i$ where $e_i$ is the $i$th elementary $k$-vector and $\vee$ is the max operator.
(b) $D^\mu K_j(x)$ is absolutely integrable and has Fourier transform

$$\Psi_\mu(r) = (2\pi)^k \int \exp(ir'x) D^\mu K_j(x)\,dx$$

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

that satisfies $\int(1+\|r\|)\sup_{b\geq 1}|\Psi_\mu(br)|\,dr<\infty\ \forall\mu$ with $|\mu|=\max(l_j/2,3)$, where $i=\sqrt{-1}$.

(c) The bandwidth parameter for each kernel estimator is $a_{jN}=O(N^{-\psi_j})$ for some $\psi_j$ such that $1/[4(l_j+2)]<\psi_j<1/[4(l_j+1)]$.

Assumption 8. If $h_0(\theta_0,x)+r(h(\theta_0,x);\theta_0)=h_0(\theta^*,x)+r(h(\theta^*,x);\theta^*)$ for all $x\in\mathcal{X}$, a set with positive probability, then $\theta_0=\theta^*$.

Assumption 1 only assumes that $(Y_i,X_i)$ are independent beyond a lag of $m$ periods. Thus our estimator, unlike previous index models that assumed that $(Y_i,X_i)$ were independent, allows for some temporal dependence in the data. Assumption 2 is standard in the semiparametric literature. Assumption 3 gives moments conditions on the indices, $h_j(\theta,X_i)$, and their derivatives.

Assumptions 4 and 5 are related to the smoothness of $f_{ji}(v_j;\theta)$ and $r_j(v_j;\theta)$. Because these assumptions are not trivially satisfied we provide primitive conditions on the indices that guarantee that these assumptions hold. Following Klein and Spady (1993), assume that after a normalization the indices in equation $j$ can be written as $h_j(\theta,X_i)=Z_{ji}+v_j(\theta,W_{ji})$, where $Z_{ji}\in\mathbf{R}^{l_j}$ for $j=1,\ldots,m$. Here $Z_{ji}=(Z_{1ji},\ldots,Z_{l_j ji})$ is a vector of distinct continuous explanatory variables with unbounded support. Thus each index in each equation contains a continuous explanatory variable that is not contained in any other index of that equation. Note that this normalization affects $h(\theta,X_i)$ but not $h_0(\theta,X_i)$. As discussed subsequently, this assumption is required for identification. Let $p_{ji}(Z_{ji}|W_{ji})$ be the conditional density of $Z_{ji}$ conditional on $W_{ji}$. Lee (1995) and Klein and Spady (1993) show that the density function of the indices and the index regression function will inherit the smoothness properties of $p_{ji}(Z_{ji}|W_{ji})$. Thus for Assumptions 4 and 5 to hold it is sufficient that $p_{ji}(Z_{ji}|W_{ji})$ is continuously differentiable in $Z_{ji}$ to order $w_j>l_j+2$ on $\mathbf{R}^{l_j}$ and $\sup_{Z_{ji}\in\mathbf{R}^{l_j}}|D^\mu p_{ji}(Z_{ji}|W_{ji})|<\infty\ \forall\mu$ with $|\mu|\leq w_j$, $\forall W_{ji}$.

Assumption 6 rules out a random trimming, thereby significantly simplifying the proofs of the asymptotic properties. However, these proofs can be modified to allow for a sample dependent trimming set $\hat{X}_N$. For example, let $h_j(\theta,X_i)=Z_{ji}+v_j(\theta,W_{ji})$ as defined previously and without loss of generality assume that the support of $v_j(\theta,W_{ji})$ is a compact set. Trim the tails of the distributions of $Z_j$ using quantile statistics with $z_{j\alpha}$ being the $\alpha$th quantile vector of $Z_j$ and $z_{j\alpha}^N$ the corresponding sample $\alpha$th quantile vector. Define the sets $X=\{X_i: z_{j\alpha}\leq Z_{ji}\leq z_{j(1-\alpha)}\text{ for } j=1,\ldots,m\}$, which is a nonstochastic compact subset of $\mathbf{R}^k$, and $\hat{X}_N=\{X_i: z_{j\alpha}^N\leq Z_{ji}\leq z_{j(1-\alpha)}^N\text{ for } j=1,\ldots,m\}$, which is stochastic. To use $\hat{X}_N$ we need to show that the error made by replacing $I_X$ with $I_{\hat{X}_N}$ is small. For example, to prove consistency we need to show that the $\sup_\theta|\hat{Q}_N(\theta,I_X)-\hat{Q}_N(\theta,I_{\hat{X}_N})|$ has order $O_p(1)$. This result and also asymptotic normality can be established by showing that $\sup_i|I_X-I_{\hat{X}_N}|$ has order $O_p(N^{-1/2})$. Because $z_{j\alpha}^N$ consistently estimates $z_{j\alpha}$ and an indicator function belongs to the euclidian class defined in Pakes and Pollard (1989), we can apply the results of Lemma 2.17 in Pakes and Pollard (1989) to show that $\sup_i|I_X-I_{\hat{X}_N}|$ has order $O_p(N^{-1/2})$. Finally, Lee (1995) showed that the

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

density function of the indices $f_{ji}(h_j(\theta, X_i);\theta)$ will be bounded away from zero on $X \times \Theta$.

Assumption 7 concerns the kernel functions and the bandwidth parameters and is identical to Assumptions NP4 and NP5 in Andrews (1995). This assumption implies that higher order kernels should be used, and the order of these kernels will depend on $l_j$. Higher order kernels reduce the asymptotic bias of the regression function, but they can not be restricted to take only nonnegative values. An example of a kernel function that does satisfy Assumptions 7a and 7b is a multivariate normal kernel proposed by Bierens (1987) and used in our Monte Carlo experiment. These assumptions are sufficient to show that the nonparametric kernel estimators of the index regression functions and their derivatives are uniformly consistent with a rate of convergence of $N^{1/4}$.

Finally, Assumption 8 is an identification condition. Although the parameters of $h_0(\theta, x)$ can be identified under certain conditions, we can only identify a function of the original parameters in $h(\theta, x)$. The existence of $h_0(\theta, x)$ is not required to identify a function of the original parameters in $h(\theta, x)$. As we will discuss in Section 5, if the indices are linear then the parameters of $h(\theta, x)$ are only identified up to scale and the intercept is not identified. Using Assumptions 1–8, we can show that $\tilde{\theta}$ is $\sqrt{N}$-consistent and asymptotically normally distributed as stated in the following theorem.

THEOREM 1. *Under Assumptions* 1–8

$$\sqrt{N}(\tilde{\theta} - \theta_0) \xrightarrow{D} N(0, M^{-1}SM^{-1})$$

*as* $N \to \infty$, *where* $M = \lim_{N\to\infty} 1/N \sum_1^N E(I_X B_i' A_i B_i)$, $S = \lim_{N\to\infty} \times \mathrm{Var}(1/\sqrt{N} \sum_1^N I_X B_i' A_i \epsilon_i)$, $B_i = (\partial/\partial\theta)h_0(X_i, \theta_0) + D_\theta r(h(X_i, \theta_0);\theta_0)$, *and* $I_X = I(X_i \in X)$.

Note that $D_\theta r(h(X_i, \theta_0);\theta_0) = -E((\partial/\partial\theta)h_0(X_i, \theta_0)|h(X_i, \theta_0)) + D_h F \times (h(X_i, \theta_0))[(\partial/\partial\theta)h(X_i, \theta_0) - E((\partial/\partial\theta)h(X_i, \theta_0)|h(X_i, \theta_0))]$, but $D_\theta r(h(X_i, \theta_0);\theta_0)$ is also equal to $r^{(1)}(h(X_i, \theta_0);\theta_0)(\partial/\partial\theta)h(X_i, \theta_0) + r^{(2)}(h(X_i, \theta_0);\theta_0)$, where $r^{(z)}$ denotes the derivative of $r$ with respect to its $z$th argument and $(\partial/\partial\theta)h(X_i, \theta_0) = (\partial/\partial\theta)\mathrm{vec}(h(X_i, \theta_0))$. The dimensions of $r^{(1)}(h(X_i, \theta_0);\theta_0)$, $r^{(2)}(h(X_i, \theta_0);\theta_0)$, and $(\partial/\partial\theta)h(X_i, \theta_0)$ are $m \times (\sum_1^m l_j)$, $m \times k^*$, and $(\sum_1^m l_j) \times k^*$, respectively. Hence, $B_i$ is $m \times k^*$.

If $A_i$ is replaced with $\Omega_i^{-1}$, $M = \lim_{N\to\infty} 1/N \sum_1^N E(I_X B_{i'} \Omega_i^{-1} B_i)$ and $S = \lim_{N\to\infty} \mathrm{Var}(1/\sqrt{N} \sum_1^N I_X B_i' \Omega_i^{-1} \epsilon_i)$. In this situation, if the observations are independent and the loss in efficiency due to the trimming is ignored, $M = S$ and $M^{-1}SM^{-1} = (1/N \sum_1^N E(I_X B_i' \Omega_i^{-1} B_i))^{-1}$.[2] Thus $\Omega_i^{-1}$ is the value of $A_i$ that minimizes the asymptotic variance. A consistent estimator, $\hat{\Omega}_i^{-1}$, can be used to form a feasible semiparametric WLS. As discussed earlier, $\Omega_i^{-1}$ depends on the model studied. In most cases it will be an unknown function of the index $(\Omega^{-1}(h(X_i, \theta_0)))$, which can be estimated nonparametrically using a preliminary estimate of $\theta$ as in our Monte Carlo study $(\hat{\Omega}^{-1}(h(X_i, \tilde{\theta})))$. Using

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

standard arguments of $m$-estimation theory, replacing $\Omega^{-1}(h(X_i,\theta_0))$ with $\hat{\Omega}^{-1}(h(X_i,\tilde{\theta}))$ will not affect the asymptotic distribution if $\hat{\Omega}^{-1}(h(X_i,\tilde{\theta})) = \Omega^{-1}(h(X_i,\theta_0)) + o_p(1)$ (see Newey, 1991).

A consistent estimator of $M$ is $\hat{M} = 1/N \sum_1^N I_X \hat{B}_i' A_i \hat{B}_i$, and because $\{I_X B_i' A_i \epsilon_i\}$ is an $m$-dependent sequence, a consistent estimate of $S$ is

$$\hat{S} = \frac{1}{N}\sum_1^N I_X \hat{B}_i' A_i \hat{\epsilon}_i \hat{\epsilon}_i' A_i \hat{B}_i + \sum_{v=1}^{m} \frac{1}{N} \sum_{i=v+1}^{N} [I_X \hat{B}_i' A_i \hat{\epsilon}_i \hat{\epsilon}_{i-v}' A_i \hat{B}_{i-v} + I_X \hat{B}_{i-v}' A_i \hat{\epsilon}_{i-v} \hat{\epsilon}_i' A_i \hat{B}_i], \tag{7}$$

where $\hat{\epsilon}_i = Y_i - h_0(X_i,\hat{\theta}) - \hat{r}(h(X_i,\hat{\theta});\hat{\theta})$ and $\hat{B}_i = (\partial/\partial\theta)h_0(X_i,\hat{\theta}) + \hat{r}^{(1)}(h(X_i,\hat{\theta});\hat{\theta})(\partial/\partial\theta)h(X_i,\hat{\theta}) + \hat{r}^{(2)}(h(X_i,\hat{\theta});\hat{\theta})$. When observations are independent and we replace $A_i$ with $\hat{\Omega}_i^{-1}$, a consistent estimate of the covariance matrix is $1/N \sum_1^N I_X \hat{B}_i' \hat{\Omega}_i^{-1} \hat{B}_i$.

Appendix B contains the proof of Theorem 1. We prove this theorem by showing that our estimator belongs to the class of MINPIN estimators proposed by Andrews (1994a). A MINPIN estimator is any estimator that minimizes a criterion function that may depend on a preliminary infinite dimensional nuisance parameter estimator. Andrews (1994a, Theorem 1) showed that MINPIN estimators, under certain regularity conditions, are asymptotically normally distributed with a rate of convergence equal to $\sqrt{N}$. In Appendix A, we state Andrews's theorem and other relevant results used to show the regularity conditions. Among these regularity conditions are the following: the consistency of the preliminary infinite dimensional nuisance parameter estimator ($\hat{\tau}$) and the fulfillment of a stochastic equicontinuity condition.

## 5. IDENTIFICATION AND EFFICIENCY

To establish identification, one needs to determine the conditions under which Assumption 8 is satisfied. As discussed in Section 4 only a function of the parameters of $h(\theta,X)$ is identified. Let

$$h_0(\theta,X) = \begin{bmatrix} X_{01}'\alpha_1 \\ \vdots \\ X_{0m}'\alpha_m \end{bmatrix} \quad \text{and}$$

$$h(\theta,X) = \begin{bmatrix} Z_{11} + W_{11}'\beta_{11},\ldots,Z_{l_1 1} + W_{l_1 1}'\beta_{l_1 1} \\ \vdots \\ Z_{1m} + W_{1m}'\beta_{1m},\ldots,Z_{l_m m} + W_{l_m m}'\beta_{l_m m} \end{bmatrix}.$$

Here we assumed linear indices without intercepts and applied the standard scale transformation on $h(\theta,X)$ for linear index models. The following lemma gives

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

the conditions that will ensure that Assumption 8 is satisfied by the normalized linear indices in the absence of cross-equations and cross-index restrictions.[3]

LEMMA 2. *Assumption* 8 *is satisfied by the normalized linear indices, if for each equation* $(say, j)$

1. $Z_{dj}$ *is a continuous explanatory variable,* $Z_{dj} \notin W_{kj}$, $Z_{dj} \notin X_{oj}$, *and* $Z_{dj} \neq Z_{kj}$ *for* $d, k = 1, \ldots, l_j$ *and* $d \neq k$.
2. *The functions* $F_1(\cdot), \ldots, F_m(\cdot)$ *in* (1) *are differentiable.*
3. *The constant function* 1 *and the derivatives functions* $F_m^j(\cdot)$, $j = 1, \ldots, d_m$ *are not linearly dependent with probability one on* $X$, *where* $F_m^j(\cdot)$ *is the partial derivative of* $F_m(\cdot)$ *with respect to the* $j$*th argument.*

The proof follows from Lemma 3 in Ichimura and Lee (1991).

The main requirement for identification is that each index in $h_j(\theta, X)$ contains a continuous explanatory variable that is not contained in any other index of $h_j(\theta, X)$ or in $h_{j0}(\theta, X)$ for $j = 1, \ldots, m$. The coefficients associated with these explanatory variables should be nonzero so they can be normalized to unity. In the sample selection model given by (3), we need a continuous explanatory variable contained in $X_2$ but not included in $X_1$ or $X_3$ and a continuous explanatory variable contained in $X_3$ but not included in $X_1$ or $X_2$ to achieve identification. Note that when $\alpha_j = 0$ for $j = 1, \ldots, m$, identification of the remaining parameters still is possible from the previous lemma.

Newey and Stoker (1993) derived a semiparametric efficiency bound for single equation index models. For the particular case of only one equation $(m = 1)$ and $\Omega^{-1} = \mathrm{var}(u|x)^{-1}$, our estimator does not achieve this semiparametric efficiency bound as a result of the trimming. Our estimator will achieve the bound if the loss in efficiency caused by the trimming is ignored. We are not aware of any semiparametric efficiency bound for MEMI models in the literature. However, for the multinomial choice model, Lee (1995) derived a semiparametric efficiency bound under index restrictions. Again our estimator will achieve this bound if the loss in efficiency due to the trimming is ignored. It is important to note that these bounds have been derived under the index assumption. As Thompson (1993) showed for the multinomial choice model, it is possible to find a semiparametric efficiency bound under the assumption of independence of the error terms and the $X$'s that is smaller than the one derived under the index assumptions.

## 6. MONTE CARLO SIMULATIONS

To analyze the finite-sample performance of the semiparametric-WLS estimator, we perform a Monte Carlo experiment with 1,000 replications. The data are simulated by the following sample selection model:

$$\begin{aligned} Y_{1i}^* &= x_{1i}\beta_{11} + u_{1i}, \\ Y_{2i}^* &= x_{2i}\beta_{21} + w_i\beta_{22} + u_{2i}, \\ Y_{3i}^* &= x_{3i}\beta_{31} + z_i\beta_{32} + u_{3i}. \end{aligned} \tag{8}$$

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

We do not observe $Y_{1i}^*$, $Y_{2i}^*$, or $Y_{3i}^*$; instead we observe $Y_{1i} = Y_{1i}^*$ and $Y_{2i} = Y_{2i}^*$ if and only if $Y_{2i}^* > 0$ and $Y_{3i}^* > 0$. The observed sample sizes are 150, 300, and 600 and are the result of 44% truncation. As discussed in Section 2, this model can be seen as the following MEMI model with two equations and two indices:

$$\begin{aligned} Y_1 &= x_{1i}\beta_{11} + F_1(x_{2i} + w_i\beta_{22}, x_{3i} + z_i\beta_{32}) + e_1 \\ Y_2 &= F_2(x_{2i} + w_i\beta_{22}, x_{3i} + z_i\beta_{32}) + e_2 \end{aligned} \tag{9}$$

The explanatory variables are generated as follows: $x_1$, $x_2$, and $x_3$ by independent logistic variables; $w$ and $z$ by independent Poisson variables with mean 2 and truncated at 5. We trim the lower tail of the empirical distributions of $x_2$ and $x_3$ by 2% and the upper tails by 2%. Because $x_2$ and $x_3$ are independent, a total of 6.4% of the observations was trimmed. We consider two different specifications of the error terms; in the first the error terms ($u$) are generated by a trivariate normal distribution $N(\mathbf{0};\mathbf{1};0.5,0.5,0)$, and in the second the error terms follow a mixture of two trivariate normal distributions $(0.75)N(\mathbf{0.5};\mathbf{1};\mathbf{0.5}) + (0.25)N(-\mathbf{1.5};\mathbf{4};\mathbf{2})$. This mixture normal has a mean = 0, variance = 2.5, covariance = 1.6, skewness coefficient = −1.04, and degree of excess = 1.82. The true parameters are $\beta_{11} = 1$, $\beta_{21} = 1$, $\beta_{22} = 0.5$, $\beta_{31} = 1$, and $\beta_{32} = 0.5$. To achieve identification we normalize $\beta_{21} = \beta_{31} = 1$. Thus, we only estimate three coefficients $\beta_{11}$, $\beta_{22}$, and $\beta_{32}$.

In our experiment we use a multivariate normal-based kernel proposed by Bierens (1987):

$$\hat{K}_j(v_i) = \det(\hat{D}_j)^{-1/2}(2\pi)^{-1}\left(\frac{3}{2}\exp\left[-\frac{v_i'\hat{D}_j^{-1}v_i}{2}\right] - \frac{3}{20}\exp\left[-\frac{v_i'\hat{D}_j^{-1}v_i}{8}\right] + \frac{1}{90}\exp\left[-\frac{v_i'\hat{D}_j^{-1}v_i}{18}\right]\right) \tag{10}$$

with $\hat{D}_j = 1/N\sum_s^N(h_j(X_s,\theta) - \bar{h}_j(X_s,\theta))'(h_j(X_s,\theta) - \bar{h}_j(X_s,\theta))$ for $j = 1,2$. This kernel function satisfies Assumptions 7a and 7b. The bandwidth parameter is chosen as $a_N = cN^{-1/13}$ with $c = 8$. Because $l_j = 2$ in our example, the bandwidth satisfies Assumption 7c. Our optimization algorithm is the downhill simplex method obtained from Press, Flannery, Teukosky, and Vetterling (1986). Because our objective function has several local minima, we estimate the model with three different starting values and use the one with the smallest objective function. Our starting values are (1.2,0.2,0.3), (0,0,0), and (0.5,0.5,0.5).

We estimate two versions of the semiparametric-WLS estimator. In the first version (SWLS1) the identity matrix is used instead of $\hat{\Omega}$. In the second version (SWLS2) we estimate $\hat{\Omega}$ using $\hat{\theta}$ from SWLS1, where $\hat{\Omega}_{jj} = [\sum\hat{e}_j^2\hat{K}\times(h(X_s,\hat{\theta}))]/[\sum\hat{K}(h(X_s,\hat{\theta}))]$ for $j = 1,2$ and $\hat{\Omega}_{ij} = [\sum\hat{e}_1\hat{e}_2\hat{K}(h(X_s,\hat{\theta}))]/[\sum\hat{K}(h(X_s,\hat{\theta}))]$ for $i,j = 1,2$ and $i \neq j$.[4] We compare the two semiparametric-WLS estimates with those from three other methods, two parametric least squares (WLS1 and WLS2) and Ichimura and Lee's semiparametric least square (SLS) applied only to the first equation. By comparing the semiparametric-WLS sim-

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

ulations results with those of the parametric methods we see how much is lost by using the semiparametric WLS when the parametric method is correctly specified and what is gained by using the semiparametric WLS when the parametric method is misspecified. Comparison with Ichimura and Lee's SLS estimator reveals the gain in efficiency due to using the second equation during the estimation of the SWLS estimators.

The parametric estimators are derived by applying least squares to the following equations:

$$Y_1 = x_{1i}\beta_{11} + \frac{\sigma_{12}}{\sigma_2}\frac{\phi(l_1)}{1-\Phi(l_1)} + \frac{\sigma_{13}}{\sigma_3}\frac{\phi(l_2)}{1-\Phi(l_2)} + e_1,$$

$$Y_2 = x_{2i}\beta_{21} + w_i\beta_{22} + \sigma_2\frac{\phi(l_1)}{1-\Phi(l_1)} + e_2, \tag{11}$$

where $l_1 = -x_{2i}(\beta_{21}/\sigma_2) - w_i(\beta_{22}/\sigma_2)$, $l_2 = -x_3(\beta_{31}/\sigma_3) - z_i(\beta_{32}/\sigma_3)$, $V(e_1) = \sigma_1^2 + (\sigma_{12}^2/\sigma_2^2)(l_1\lambda(l_1) - \lambda^2(l_1)) + (\sigma_{13}^2/\sigma_3^2)(l_2\lambda(l_2) - \lambda^2(l_2))$, $V(e_2) = \sigma_2^2(1 + l_1\lambda(l_1) - \lambda^2(l_1))$, and $\text{cov}(e_1, e_2) = \sigma_{12}(1 + l_1\lambda(l_1) - \lambda^2(l_1))$, with $\lambda(l) = \phi(l)/\Phi(-l)$. Two versions of this WLS estimator are obtained; in the first version we use the identity matrix as the variance-covariance matrix (WLS1), and in the second version we use the true variance-covariance matrix (WLS2).[5]

Tables 1 and 2 report the mean, standard deviation (SD), and root mean square error (RMSE) for $\beta_{11}$, $\beta_{22}$, and $\beta_{32}$. Table 1 contains the results for the trivariate normal distribution, and Table 2 contains the results for the mixture trivariate normal. The parametric estimators will be consistent in Table 1 but not in Table 2. In examining the results from Tables 1 and 2 the following conclusions can be drawn. First, all methods estimate $\beta_{11}$ very well regardless of the distribution function of the error terms, and the RMSE's of $\beta_{11}$ are about the same for all estimators. This result is not surprising because $x_1$ is uncorrelated with the regressors in the other two indices. Second, the semiparametric estimator proposed in this paper performs very well estimating $\beta_{22}$. The efficiency loss of using the SWLS2 estimator when the model is correctly specified in Table 1 is quite small, whereas the gains in Table 2 are considerable when the parametric methods are biased. Third, none of the methods estimate $\beta_{32}$ very accurately, and the standard deviations of the SWLS1 and SWLS2 estimates of $\beta_{32}$ are very large compared to the standard deviations of $\beta_{11}$ and $\beta_{22}$.

Note that we only observe the model if $Y_2^*$ and $Y_3^*$ are greater than zero. However, although we observe the actual values of $Y_2^*$ when $Y_2^* > 0$ and $Y_3^* > 0$, we only observe that the latent variable $Y_3^*$ is positive without observing its value. Therefore, the selectivity term given by the third equation of (8) contains very little information compared to the selectivity term of the second equation. The lack of information provided by the index associated with the second selectivity equation may explain the large standard errors associated with $\beta_{32}$ for the semiparametric WLS and the estimates of $\beta_{22}$ and $\beta_{32}$ using Ichimura and Lee's es-

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

**TABLE 1.** Trivariate normal $N(\mathbf{0};\mathbf{1};0.5,0.5,0)$

| | | B11 | | | B22 | | | B32 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | $N$ | Mean | SD | RMSE | Mean | SD | RMSE | Mean | SD | RMSE |
| WLS1 | 150 | 1.0003 | 0.0445 | 0.0446 | 0.4906 | 0.0482 | 0.0491 | 0.6380 | 0.1751 | 0.2229 |
| | 300 | 0.9996 | 0.0313 | 0.0313 | 0.4937 | 0.0359 | 0.0365 | 0.6421 | 0.1329 | 0.1946 |
| | 600 | 1.0005 | 0.0219 | 0.0219 | 0.4983 | 0.0257 | 0.0257 | 0.6375 | 0.0974 | 0.1685 |
| WLS2 | 150 | 0.9994 | 0.0417 | 0.0417 | 0.4883 | 0.0509 | 0.0521 | 0.6461 | 0.2091 | 0.2551 |
| | 300 | 0.9993 | 0.0302 | 0.0302 | 0.4907 | 0.0381 | 0.0393 | 0.6431 | 0.1739 | 0.2252 |
| | 600 | 1.0001 | 0.0196 | 0.0196 | 0.4981 | 0.0289 | 0.0291 | 0.6439 | 0.1127 | 0.1828 |
| SLS | 150 | 1.0006 | 0.0456 | 0.0456 | 0.4464 | 0.6651 | 0.6672 | 0.4997 | 0.6623 | 0.6623 |
| | 300 | 0.9997 | 0.0319 | 0.0319 | 0.5511 | 0.5635 | 0.5659 | 0.5783 | 0.5668 | 0.5731 |
| | 600 | 1.0007 | 0.0223 | 0.0224 | 0.5951 | 0.3904 | 0.4018 | 0.5842 | 0.4246 | 0.4329 |
| SWLS1 | 150 | 1.0005 | 0.0456 | 0.0456 | 0.4902 | 0.0694 | 0.0701 | 0.4414 | 0.7946 | 0.7968 |
| | 300 | 0.9997 | 0.0319 | 0.0319 | 0.4884 | 0.0495 | 0.0508 | 0.5281 | 0.7010 | 0.7016 |
| | 600 | 1.0008 | 0.0221 | 0.0221 | 0.4875 | 0.0341 | 0.0363 | 0.5823 | 0.4865 | 0.4896 |
| SWLS2 | 150 | 1.0002 | 0.0447 | 0.0446 | 0.4924 | 0.0599 | 0.0603 | 0.5087 | 0.7256 | 0.7257 |
| | 300 | 0.9995 | 0.0311 | 0.0311 | 0.4913 | 0.0430 | 0.0439 | 0.5882 | 0.5915 | 0.5980 |
| | 600 | 1.0006 | 0.0217 | 0.0217 | 0.4901 | 0.0299 | 0.0315 | 0.5908 | 0.4125 | 0.4223 |

*Note:* Statistics are for 1,000 replications. $N$ = sample size, SD = standard deviation, RMSE = root mean square error, WLS1 = parametric WLS using identity matrix, WLS2 = parametric WLS using optimal weighting matrix, SLS = Ichimura and Lee's semiparametric least square, SWLS1 = semiparametric WLS using identity matrix, and SWLS2 = semiparametric WLS using optimal weighting matrix.

**TABLE 2.** Mixture trivariate normal $(0.75)N(0.5;1;0.5) + (0.25)N(-1.5;4;2)$

| | | B11 | | | B22 | | | B32 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | $N$ | Mean | SD | RMSE | Mean | SD | RMSE | Mean | SD | RMSE |
| WLS1 | 150 | 1.0002 | 0.0601 | 0.0601 | 0.5669 | 0.0717 | 0.0981 | 0.6216 | 0.2213 | 0.2525 |
| | 300 | 0.9988 | 0.0445 | 0.0445 | 0.5637 | 0.0604 | 0.0878 | 0.6186 | 0.2060 | 0.2377 |
| | 600 | 0.9994 | 0.0346 | 0.0346 | 0.5648 | 0.0507 | 0.0823 | 0.6228 | 0.1936 | 0.2293 |
| WLS2 | 150 | 0.9999 | 0.0555 | 0.0555 | 0.5609 | 0.0669 | 0.0905 | 0.6069 | 0.1676 | 0.1988 |
| | 300 | 0.9988 | 0.0411 | 0.0411 | 0.5597 | 0.0581 | 0.0833 | 0.6032 | 0.2034 | 0.2281 |
| | 600 | 0.9998 | 0.0326 | 0.0326 | 0.5586 | 0.0492 | 0.0765 | 0.6081 | 0.1897 | 0.2184 |
| SLS | 150 | 0.9995 | 0.0536 | 0.0536 | 0.4012 | 0.6947 | 0.7015 | 0.4399 | 0.6701 | 0.6727 |
| | 300 | 0.9998 | 0.0372 | 0.0372 | 0.5341 | 0.5976 | 0.5986 | 0.4911 | 0.5831 | 0.5831 |
| | 600 | 1.0003 | 0.0267 | 0.0267 | 0.5876 | 0.4332 | 0.4405 | 0.5566 | 0.4362 | 0.4399 |
| SWLS1 | 150 | 0.9994 | 0.0536 | 0.0536 | 0.4928 | 0.0815 | 0.0816 | 0.4194 | 0.7227 | 0.7272 |
| | 300 | 0.9998 | 0.0372 | 0.0372 | 0.4927 | 0.0582 | 0.0587 | 0.4551 | 0.6535 | 0.6551 |
| | 600 | 1.0003 | 0.0267 | 0.0267 | 0.4919 | 0.0408 | 0.0416 | 0.5530 | 0.5025 | 0.5053 |
| SWLS2 | 150 | 0.9995 | 0.0530 | 0.0530 | 0.4951 | 0.0733 | 0.0725 | 0.4356 | 0.7295 | 0.7323 |
| | 300 | 1.0001 | 0.0362 | 0.0362 | 0.4947 | 0.0514 | 0.0516 | 0.4788 | 0.6082 | 0.6086 |
| | 600 | 1.0005 | 0.0259 | 0.0259 | 0.4933 | 0.0369 | 0.0375 | 0.5585 | 0.4426 | 0.4464 |

*Note:* Statistics are for 1,000 replications. $N$ = sample size, SD = standard deviation, RMSE = root mean square error, WLS1 = parametric WLS using identity matrix, WLS2 = parametric WLS using optimal weighting matrix, SLS = Ichimura and Lee's semiparametric least square, SWLS1 = semiparametric WLS using identity matrix, and SWLS2 = semiparametric WLS using optimal weighting matrix.

**Table 3.** SWLS1 with different bandwidth factors

| | B11 | | | B22 | | | B32 | | |
|---|---|---|---|---|---|---|---|---|---|
| Factor | Mean | SD | RMSE | Mean | SD | RMSE | Mean | SD | RMSE |
| *Normal distribution, sample size 300, 500 replications* | | | | | | | | | |
| 1 | 1.0011 | 0.0367 | 0.0368 | 0.3312 | 0.2102 | 0.2695 | 0.1792 | 0.5131 | 0.6051 |
| 3 | 0.9998 | 0.0335 | 0.0335 | 0.4839 | 0.0531 | 0.0554 | 0.4156 | 1.3564 | 1.3591 |
| 6 | 0.9997 | 0.0337 | 0.0337 | 0.4868 | 0.0514 | 0.0531 | 0.5211 | 1.0094 | 1.0096 |
| 9 | 0.9996 | 0.0338 | 0.0338 | 0.4884 | 0.0511 | 0.0524 | 0.5099 | 0.5758 | 0.5759 |
| 12 | 0.9996 | 0.0337 | 0.0337 | 0.3882 | 0.1234 | 0.1664 | 0.2171 | 0.1702 | 0.3302 |
| *Mixture distribution, sample size 300, 500 replications* | | | | | | | | | |
| 1 | 1.0006 | 0.0391 | 0.0391 | 0.3388 | 0.2133 | 0.2674 | 0.1819 | 0.5201 | 0.6095 |
| 3 | 1.0013 | 0.0364 | 0.0364 | 0.4898 | 0.0661 | 0.0668 | 0.4517 | 1.1472 | 1.1482 |
| 6 | 1.0013 | 0.0364 | 0.0364 | 0.4922 | 0.0611 | 0.0615 | 0.4779 | 0.8625 | 0.8628 |
| 9 | 1.0013 | 0.0363 | 0.0364 | 0.4931 | 0.0605 | 0.0608 | 0.4233 | 0.5341 | 0.5395 |
| 12 | 1.0116 | 0.0462 | 0.0477 | 0.2723 | 0.1315 | 0.2629 | 0.2001 | 0.2061 | 0.3639 |

timator when both $Y_2^*$ and $Y_3^*$ are latent. It is worth noting that although the semiparametric-WLS estimates of $\beta_{32}$ appear to be unreliable, the estimates of $\beta_{11}$ and $\beta_{22}$ remain unaffected. This is consistent with the conclusions of Maddala (1983, p. 267) when discussing parametric estimation of selectivity bias equations caused by a latent truncated variable. Parametric estimators also perform poorly, even when the model is correctly specified.

The choice of 8 as the constant in our bandwidth ($cN^{-1/13}$) was arbitrary because any value of $c$ will also satisfy Assumption 7c. Table 3 reports the SWLS1 estimates with different values of $c$ for the trivariate normal and the mixture normal with 300 observations. Whereas the estimates of $\beta_{32}$ are very sensitive to the bandwidth parameter, the estimates of $\beta_{11}$ and $\beta_{22}$ appear to be fairly similar for a range of the bandwidth parameter (between 3 and 9). In practical applications it would be useful to have a rule to select this constant. Unfortunately, neither the theoretical results from Section 4 nor our Monte Carlo study provide a guideline for selecting $c$.[6] This is a topic for further research.

Each replication of the SWLS1 took approximately 30.52 seconds of cpu time for 150 observations and 388.41 seconds of cpu time for 600 observations running in a Pentium 200 Pro using Fortran 90. For comparison, the SLS took roughly the same time, but the WLS1 took only 0.28 seconds and 1.39 seconds of cpu time with 150 and 700 observations, respectively.

## 7. CONCLUSION

This paper proposes a semiparametric estimator for multiple equations multiple index (MEMI) models, derives the variance-covariance matrix of the estimator, and shows that the estimator satisfies the standard desirable asymptotic properties of $\sqrt{N}$-consistency and asymptotic normality. It also discusses the identification of the model and examines the finite-sample behavior in a Monte Carlo experiment.

Examples of MEMI models include sample selection models with multiple equation and selection terms, Dubin and McFadden's utility maximizing models with discrete and continuous choices, and the multinomial choice model. The proposed estimator can be used to estimate these important econometric models without specifying a parametric distribution function for the error terms. Considering that the distribution function of the error terms is usually unknown, an estimator that relaxes this assumption can be potentially very useful. On the other hand, the estimator proposed in this paper, like most semiparametric estimators, is considerably more expensive to compute compared to a well-behaved parametric estimator. As a result of this cost, in practice the proposed estimator may best be used in concurrence with parametric methods.

### *NOTES*

1. Lee (1995) proposed a semiparametric maximum likelihood estimator, but this estimator can only be used to estimate the multinomial choice model.

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

2. If the observations are independent, then $S = 1/N \sum E[(I_X B_i - E(I_X B_i|h(X_i,\theta_0))) \times \Omega_i^{-1}(I_X B_i - E(I_X B_i|h(X_i,\theta_0)))]$. Although $E(B_i|h(X_i,\theta_0)) = 0$, $E(I_X B_i|h(X_i,\theta_0)) \neq 0$.

3. Cross-equation and cross-index restrictions in general will help identification.

4. For the semiparametric estimators we imposed a penalty when $\hat{f}(v_i,\theta) < 0.001$.

5. The WLS estimates $\beta_{21}$, $\beta_{31}$, $\sigma_{12}$, $\sigma_{13}$, $\sigma_2$ in addition to $\beta_{11}$, $\beta_{22}$, and $\beta_{32}$.

6. A reasonable candidate for $c$ could be the constant from an optimal bandwidth for estimating a multivariate density of the indices (see Scott, 1992). Note that this bandwidth will depend on the order of the kernel and the true distribution of the indices. Thus in practice we need to have an idea of the shape of the distribution of the indices. For a single index SLS estimator Cavanagh and Sherman (1998) follow this approach. Calculating this constant for multiple indices would be a very challenging problem, and we did not calculate this constant for our Monte Carlo simulations.

## *REFERENCES*


Andrews, D.W.K. (1992) Generic uniform convergence. *Econometric Theory* 8, 241–257.

Andrews, D.W.K. (1994a) Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* 62, 43–72.

Andrews, D.W.K. (1994b) Empirical process methods in econometrics. In R.F. Engle & D.L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2247–2294. New York: North-Holland.

Andrews, D.W.K. (1995) Nonparametric kernel estimation for semiparametric models. *Econometric Theory* 11, 560–596.

Bierens, H.J. (1987) Kernel estimators of regression functions. In T.F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, vol. I, pp. 99–144. New York: Cambridge University Press.

Cavanagh, C. & R.P. Sherman (1998) Rank estimator for monotonic index models. *Journal of Econometrics* 84, 351–381.

Dubin, J. & D.L. McFadden (1984) An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52, 345–362.

Gallant, A.R. (1987) *Nonlinear Statistical Models.* New York: Wiley.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71–120.

Ichimura, H. & L.-F. Lee (1991) Semiparametric estimation of multiple index models. In W.A. Barnett, J.L. Powell, & G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth Symposium in Economic Theory and Econometrics.* New York: Cambridge University Press.

Klein, R.W. & R.H. Spady (1993) An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–421.

Lee, L.-F. (1995) Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics* 65, 381–428.

Maddala, G. (1983) *Limited Dependent and Qualitative Variables in Econometrics.* New York: Cambridge University Press.

Newey, W.K. (1991) The Asymptotic Variance of Semiparametric Estimators. Working paper 583, Department of Economics, MIT.

Newey, W.K. & T.M. Stoker (1993) Efficiency of weighted average derivative estimators and index models. *Econometrica* 61, 1199–1223.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Press, W.H., B.P. Flannery, S.A. Teukosky, & W.T. Vetterling (1986) *Numerical Recipes in Fortran.* New York: Cambridge University Press.

Scott, D.W. (1992) *Multivariate Density Estimation.* New York: Wiley.

Thompson, T.S. (1993) Some efficiency bounds for semiparametric discrete choice models. *Journal of Econometrics* 58, 257–274.

# APPENDIX A

*MINPIN Estimators.* Andrews (1994a) defined a sequence of MINPIN estimators $\{\hat{\theta}\}$ as any sequence of random variables such that

$$d(\bar{m}_N(\hat{\theta},\hat{\tau}),\hat{\gamma}) = \inf_{\theta\in\Theta} d(\bar{m}_N(\theta,\hat{\tau}),\hat{\gamma}) \qquad wp \to 1,$$

where $\bar{m}_N(\theta,\tau) = (1/N)\sum_1^N m_i(\theta,\tau)$ and $m_i(\theta,\tau)$ denotes $m_i(W_i,\theta,\tau)$, a function from $\mathbf{R}^k \times \Theta \times \Upsilon$ to $\mathbf{R}^v$, $\Theta \subset \mathbf{R}^p$, $\hat{\tau}$ is a random element of $\Upsilon$ $wp \to 1$, $\hat{\gamma}$ is a random element of $\Gamma$ (and $\hat{\tau}$ and $\hat{\gamma}$ depend on $N$ in general), $\Upsilon$ and $\Gamma$ are pseudometric spaces, and $d(\cdot,\cdot)$ is a nonrandom, real-valued function.

Assume $d(m,\gamma) = \frac{1}{2}m'm$ and define an empirical process $v_N(\cdot)$ by

$$\nu_N(\tau) = \sqrt{N}(\bar{m}_N(\theta_0,\tau) - \bar{m}_N^*(\theta_0,\tau)) \quad \text{for } \tau \in \Upsilon,$$

where $\bar{m}_N^*(\theta_0,\tau) = 1/N\sum_1^N Em_i(\theta_0,\tau)$.

**Assumption A.**

(a) $\hat{\theta} \xrightarrow{p} \theta_0 \in \Theta \subset \mathbf{R}^p$ and $\theta_0$ is in the interior of $\Theta$.
(b) $P(\hat{\tau} \in \Upsilon) \to 1$, $\hat{\tau} \xrightarrow{p} \tau_0$ for some $\tau_0 \in \Upsilon$.
(c) $\sqrt{N}\bar{m}_N^*(\theta_0,\hat{\tau}) \xrightarrow{p} 0$.
(d) $\nu_N(\tau_0) \xrightarrow{d} N(0,S)$.
(e) $\{\nu_N(\cdot)\}$ is stochastically equicontinuous at $\tau_0$.
(f) $m_i(\theta,\tau)$ is continuously differentiable in $\theta$ on $\Theta$, $\forall\tau \in \Upsilon$, $\forall i \geq 1$, $\forall\omega \in \Omega$. Here $\{m_i(\theta,\tau)\}$ and $\{(\partial/\partial\theta')m_i(\theta,\tau)\}$ satisfy uniform WLLN over $\Theta \times \Upsilon$. The expressions $m(\theta,\tau) = \lim_{N\to\infty}(1/N)\sum_1^N Em_i(\theta,\tau)$ and

$$M(\theta,\tau) = \lim_{N\to\infty}\frac{1}{N}\sum_1^N E\left(\frac{\partial}{\partial\theta'}\right)m_i(\theta,\tau)$$

each exist uniformly over $\Theta \times \Upsilon$ and are continuous at $(\theta_0,\tau_0)$ with respect to some pseudometric on $\Theta \times \Upsilon$ for which $(\hat{\theta},\hat{\tau}) \xrightarrow{p} (\theta_0,\tau_0)$.

THEOREM 3. *Under Assumption* A *every sequence of* MINPIN *estimators* $\{\hat{\theta}\}$ *satisfies*

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, M^{-1}SM^{-1}),$$

*where* $M = M(\theta_0,\tau_0)$ *and* $S = \mathrm{Var}(\sqrt{N}\bar{m}_N(\theta_0,\tau_0))$.

**Proof.** See Andrews (1994a).

*Stochastic Equicontinuity.* Let $\{W_t\}$ be a sequence of $\mathcal{W}$-valued rv's where $\mathcal{W} \subset R^k$. Let $\mathcal{T}$ be a pseudometric space with pseudometric $\rho$. Let $\mathcal{M} = \{m(\cdot,\tau):\tau \in \mathcal{T}\}$ be a class of $R^v$-valued functions defined on $\mathcal{W}$ and indexed by $\tau \in \mathcal{T}$. Define an empirical process $v_T(\tau)$ by

$$\nu_T(\tau) = \frac{1}{\sqrt{T}}\sum_1^T (m(W_t,\tau) - Em(W_t,\tau)) \quad \text{for } \tau \in \Upsilon.$$

We will say that a sequence of empirical processes $\{v_T(\cdot):T \geq 1\}$ is stochastic equicontinuous if $v_T(\cdot)$ is continuous in $\tau$ uniformly over $\mathcal{T}$ at least with high probability and for

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

$T$ large. Primitive conditions for stochastic equicontinuity are given in the following theorem from Andrews (1994b).

THEOREM 4. *$\{v_T(\cdot): T \geq 1\}$ will be stochastic equicontinuous with pseudometric*

$$\rho_Y(\tau_1, \tau_2) = \sup_{N^* \geq 1; i \leq N^*} [E(m_i(\theta_0, \hat{\tau}) - m_i(\theta_0, \tau_0))^2]^{1/2}$$

*under the following assumptions*:

(a) *$\{W_t\}$ is an m-dependent sequence of rv's.*
(b) *$\lim_{T\to\infty}(1/T)\sum_1^T E\overline{M}^{2+\delta}(W_t) < \infty$ for some $\delta > 0$ where $\overline{M}$ is a real function on $\mathcal{W}$ for which $|m(\cdot)| \leq \overline{M}(\cdot)\ \forall m \in \mathcal{M}$.*
(c) *$\mathcal{M}$ satisfies Ossiander's $L^p$ entropy condition with $p = 2$ and has envelope $\overline{M}$.*

As shown in Andrews (1994b), several classes of functions satisfy Ossiander's entropy conditions, and functions from these classes can be mixed and matched to obtain more general results. One of the classes that satisfy Ossiander's entropy conditions is a type V class of functions defined as follows.

DEFINITION 5. *A class $\mathcal{T}$ of real functions on $\mathcal{W}$ is called a type V class under* **P** *with index $p \in [2,\infty]$ if*

(i) *each $\tau \in \mathcal{T}$ depends on w only through a subvector $w_a$ of dimension $k_a \leq k$,*
(ii) *$\mathcal{W}_a^*$ is such that $\mathcal{W}_a^* \cap \{w_a \in R^{k_a}: \|w_a\| \leq r\}$ is a connected compact set $\forall r > 0$,*
(iii) *for some real number $q > k_a/2$, each $\tau \in \mathcal{T}$ has partial derivatives of order $[q]$ on $\mathcal{W}$, the $[q]$th-order partial derivatives of $\tau$ satisfy a Lipschitz condition with exponent q-$[q]$ and some Lipschitz constant $C_q$ that does not depend on $\tau$, and $\mathcal{W}_a^*$ is a convex set.*
(iv) *$\sup_{t\leq T, T\geq 1} E\|W_{at}\|^{\zeta} < \infty$ for some $\zeta > pqk_a/(2q - k_a)$ under* **P**.

If $\mathcal{W}_a^* = R^{k_a}$, the preceding condition (ii) holds.

*Consistency*

LEMMA 6. *Under Assumptions* 1–8, $\hat{\theta} \to \theta_0$.

**Proof.** Let $Q(\theta) = Q(\theta, \tau_0)$, where $Q(\theta, \tau_0)$ is defined as

$$Q(\theta, \tau_0) = \lim_{N\to\infty} \frac{1}{N} \sum_1^N E[I_X(Y_i - h_0(\theta, X_i) - r(h(\theta, X_i); \theta))' \times A_i(Y_i - h_0(\theta, X_i) - r(h(\theta, X_i); \theta))]$$

and

$$Q(\theta, \hat{\tau}) = \lim_{N\to\infty} \frac{1}{N} \sum_1^N E[I_X(Y_i - h_0(\theta, X_i) - \hat{r}(h(\theta, X_i); \theta))' \times A_i(Y_i - h_0(\theta, X_i) - \hat{r}(h(\theta, X_i); \theta))].$$

By Lemma A.1 of Andrews (1994a) to ensure consistency of the semiparametric WLS, it is sufficient to show

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

(a) $\sup_{\theta\in\Theta}|\hat{Q}_N(\theta) - Q(\theta)| \xrightarrow{p} 0$, and
(b) $\inf_{\theta\in\Theta/\Theta^*} Q(\theta) > Q(\theta_0)$.

Using the triangle inequality:

$$\sup_{\theta\in\Theta}|\hat{Q}_N(\theta,\hat{\tau}) - Q(\theta,\tau_0)| \leq \sup_{\theta\in\Theta}|\hat{Q}_N(\theta,\hat{\tau}) - Q(\theta,\hat{\tau})| + \sup_{\theta\in\Theta}|Q(\theta,\hat{\tau}) - Q(\theta,\tau_0)|. \quad \textbf{(A.1)}$$

Applying the generic uniform weak law of large numbers over $\Theta$ given in Andrews (1992, Theorem 3) to

$$Q_i(\theta,\hat{\tau}) = I_{X_N}(Y_i - h_0(\theta,X_i) - \hat{r}(h(\theta,X_i);\theta))'A_i(Y_i - h_0(\theta,X_i) - \hat{r}(h(\theta,X_i);\theta)),$$

we conclude that the first term of the inequality given in (A.1) converges to zero. To show that the second term in the inequality also converges to zero, note that after some algebra and using standard inequalities, it can be shown that

$$\sup_{\theta\in\Theta}|Q(\theta,\hat{\tau}) - Q(\theta,\tau_0)| \leq \sup_{\theta\in\Theta} A_1(\theta,\tau_0,\hat{\tau}) + \sup_{\theta\in\Theta} A_2(\theta,\tau_0,\hat{\tau}),$$

where $A_1(\theta,\tau_0,\hat{\tau}) = \lim_{N\to\infty} 1/N \sum_1^N E|I_X[\hat{r}(h(\theta,X_i);\theta) - r(h(\theta,X_i);\theta)]'A_i \times [\hat{r}(h(\theta,X_i);\theta) + r(h(\theta,X_i);\theta)]|$ and $A_2(\theta,\tau_0,\hat{\tau}) = \lim_{N\to\infty} 2(1/N) \sum_1^N E|I_X[\hat{r} \times (h(\theta,X_i);\theta) - r(h(\theta,X_i);\theta)]'A_i(Y_i - h_0(\theta,X_i))|$.

Using Holder's inequality and the triangle inequality, combined with the fact that uniform convergence implies $L^Q$ convergence, we conclude that to establish $\sup_\Theta A_z(\theta,\tau_0,\hat{\tau}) \xrightarrow{p} 0$ $(z = 1,2)$, it is sufficient to show that

$$\sup_{\Theta\times X}|[\hat{r}_j(h_j(\theta,X_i);\theta) - r_j(h_j(\theta,X_i);\theta)]| \xrightarrow{p} 0 \quad \text{for } j = 1,\ldots,m.$$

In Lemma 7 we show that $\sup_{\Theta\times V_j^*\cap V_j}|\hat{r}_j(v_j;\theta) - r_j(v_j;\theta)| \xrightarrow{p} 0$, where $V_j^*$ is a bounded subset of $\mathbf{R}^{l_j}$ and $V_j = \{v_j : \inf_{\theta\in\Theta} 1/N \sum_1^N f_{ji}(v_j;\theta) \geq d\}$. Let $V_j(X,\theta) = \{h_j(\theta,X_i) : X_i \in X \text{ and } \theta \in \Theta\}$. By Assumption 6 there exists a bounded subset of $\mathbf{R}^{l_j}$ that includes $V_j(X,\theta)$. Assumption 6 also guarantees that $\inf_{\Theta\times X} f_{ji}(h_j(\theta,X_i);\theta) > 0$. Therefore $V_j(X,\theta) \subset V_j^* \cap V_j$, and the results of Lemma 7 establish that $\sup_\Theta A_z(\theta,\tau_0,\hat{\tau}) \xrightarrow{p} 0$ $(z = 1,2)$.

Let $\xi_i = Y_i - h_0(\theta_0,X_i)$ $r(h(\theta_0,X_i);\theta_0)$, then $Q(\theta_0) = \lim_{N\to\infty} 1/N \sum_1^N EI_X \xi_i' A_i \xi_i$ and $Q(\theta) = \lim_{N\to\infty} 1/N \sum_1^N EI_X \xi_i' A_i \xi_i + z(\theta)$, where

$$\begin{aligned} z(\theta) = \lim_{N\to\infty} \frac{1}{N} \sum_1^N E[I_X(h_0(\theta_0,X_i) - h_0(\theta,X_i) + r(h(\theta_0,X_i);\theta_0) \\ - r(h(\theta,X_i);\theta))'A_i(h_0(\theta_0,X_i) - h_0(\theta,X_i) \\ + r(h(\theta_0,X_i);\theta_0) - r(h(\theta,X_i);\theta))], \end{aligned}$$

because $A_i$ is a positive semidefinite matrix, then $z(\theta) \geq 0$ and $Q(\theta) - Q(\theta_0) = z(\theta) \geq 0$. Note that $z(\theta_0) = 0$ and $Q(\theta)$ achieve a minimum at $\theta_0$. Finally, Assumption 8 ensures that $\theta_0$ is the only value of $\theta$ that minimizes $Q(\theta)$. ■

LEMMA 7. *Under Assumptions* 1–7 *as* $N \to \infty$

$$\sup_{\Theta\times V_j^*\cap V_j}|D^{\mu_v}D^{\mu_\theta}\hat{r}_j(v_j;\theta) - D^{\mu_v}D^{\mu_\theta}r_j(v_j;\theta)| \xrightarrow{p} 0 \quad \textbf{(A.2)}$$

*for* $|\mu_v| \leq 2$ *and* $|\mu_\theta| \leq 2$.

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

$$N^{1/4} \sup_{V_j^* \cap V_j} |r_j(v_j;\theta_0) - \hat{r}_j(v_j;\theta_0)| \xrightarrow{p} 0 \tag{A.3}$$

$$N^{1/4} \sup_{V_j \cap V_j^*} |r_j^{(z)}(v_j;\theta_0) - \hat{r}_j^{(z)}(v_j;\theta_0)| \xrightarrow{p} 0 \tag{A.4}$$

*for* $j = 1,\ldots,l$ *and* $z = 1,2$, *where* $V_j^*$ *is a bounded subset* $\mathbf{R}^{l_j}$, $V_j = \{v_j: \inf_{\theta\in\Theta} 1/N \sum_1^N f_{ji}(v_j;\theta) \geq d\}$, *and* $d > 0$. *The term* $r_j^{(z)}$ *denotes the derivative of* $r_j$ *with respect to its zth argument,* $D^{\mu_v} r_j(v;\theta)$ *denotes the partial derivative with respect to* $v$ *of order* $\mu_v$, *and* $D^{\mu_\theta} r_j(v;\theta)$ *denotes the partial derivative with respect to* $\theta$ *of order* $\mu_\theta$.

**Proof.** We can establish (A.2) for each equation $j$, $j = 1,\ldots,m$ by verifying Assumptions NP1*-A and NP2*–NP5* of Theorem 3(b) of Andrews (1995) with $|\lambda_j| = 0,1,2$, $q_j = 2$, $k_j = l_j$, and $w_j > l_j/2$ provided that $a_N^j = O(N^{-\psi_j})$ with $\psi_j < 1/(3l_j + 4)$. NP1*-A holds by Assumptions 1 and 3. NP2* holds by Assumption 4. NP3* holds by Assumption 5. NP4* and NP5* hold by Assumption 7.

For equation $j$, (A.3) can be established by applying Theorem 1(b) of Andrews (1995) with $k_j = l_j$, $|\lambda_j| = 0$, $\eta_j = \infty$, $a_N^j = O(N^{-\psi_j})$, $w_j = l_j + 3$, and $d_N = d$ to obtain

$$N^{\mathcal{N}} \sup_{V_j^* \cap V_j} |r_j(v_j;\theta_0) - \hat{r}_j(v_j;\theta_0)|$$
$$= O_p(N^{-1/2+\mathcal{N}+\psi l_j}) + O_p(N^{\mathcal{N}-(l_j+3)\psi}). \tag{A.5}$$

Then to obtain $N^{1/4}$ consistency we need $1/(4(l_j + 3)) < \psi_j < 1/4l_j$, which is implied by Assumption 7.

Let $r_{jd}^{(1)}(v_{ji};\theta)$ be an element of $r^{(1)}(v_i;\theta)$. By Assumptions 1–7 and applying Theorem 1(b) in Andrews (1995) with $k_j = l_j$, $|\lambda_j| = 1$, $\eta_j = \infty$, $a_N^j = O(N^{-\psi_j})$, $w_j = l_j + 3$, and $d_N = d$,

$$N^{\mathcal{N}} \sup_{\{v_j: 1/N \sum_1^N f_{ji}(v_j;\theta_0) \geq d\}} |\hat{r}_{jd}^{(1)}(v_j;\theta_0) - r_{jd}^{(1)}(v_j;\theta_0)|$$
$$= O_p(N^{-1/2+\mathcal{N}+\psi(l_j+1)}) + O_p(N^{-\psi(l_j+2)}). \tag{A.6}$$

Because by Assumption 7 $1/(4(l_j + 2)) < \psi < 1/(4(l_j + 1))$, then

$$N^{1/4} \sup_{\{v_j: 1/N \sum_1^N f_{ji}(v_j;\theta_0) \geq d\}} |\hat{r}_{jd}^{(1)}(v_j;\theta_0) - r_{jd}^{(1)}(v_j;\theta_0)| \xrightarrow{p} 0.$$

Thus, each element of $|\hat{r}^{(1)}(v;\theta_0) - r^{(1)}(v;\theta_0)|$ converges to zero uniformly over $v_j \in V_j \cap V_j^*$ with a rate of convergence of $N^{1/4}$. As discussed in Andrews (1995, pp. 26–28) we can show that

$$N^{1/4} \sup_{\{v_j: 1/N \sum_1^N f_{ji}(v_j;\theta_0) \geq d\}} |\hat{r}_{jd}^{(2)}(v_j;\theta_0) - r_{jd}^{(2)}(v_j;\theta_0)| \xrightarrow{p} 0 \tag{A.7}$$

using Theorem 1 and Lemma A1 of his paper. Because Assumptions 1–7 are sufficient to ensure the use of these results, (A.4) holds. ■

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

# APPENDIX B

**Proof of Theorem 1.** The semiparametric WLS belongs to the class of MINPIN estimators. We can show this by defining $d(\bar{m}_N(\theta,\hat{\tau}),\gamma) = \frac{1}{2}\bar{m}_N(\theta,\hat{\tau})'\bar{m}_N(\theta,\hat{\tau})$ where $\bar{m}_N(\theta,\tau) = (1/N)\sum_1^N m_i(\theta,\tau)$ and

$$m_i(\theta,\hat{\tau}) = I_{X_N}\left(\frac{\partial}{\partial\theta}h_0(\theta,X_i) + D_\theta\hat{r}(h(X_i,\theta);\theta)\right)'A_i(Y_i - h_0(\theta,X_i) - \hat{r}(h(X_i,\theta);\theta)).$$

From the first-order condition of the minimization problem given by (6), $\bar{m}_N(\hat{\theta},\hat{\tau}) = 0$. Thus, $\bar{m}_N(\theta,\hat{\tau})'\bar{m}_N(\theta,\hat{\tau})$ is minimized at $\hat{\theta}$. The expression $\hat{\tau}(v,\theta) = \hat{r}(v;\theta)$ is the nonparametric estimator of $\tau_0(v,\theta) = r(v;\theta)$. The term $\Upsilon$ is a space of functions to which $\tau = (\tau_1,\ldots,\tau_m)$ belongs, such that $\tau_j(v,\theta)$ is an $\mathbf{R}^{l_j}\times\Theta\to\mathbf{R}$ function and $\tau_j(v_j,\theta)$, $\tau_j^{(1)}(v_j,\theta)$, and $\tau_j^{(2)}(v_j,\theta)$ are continuously differentiable in $v_j$ to order $q > l_j/2$. Here $M = \lim_{N\to\infty} 1/N\sum_1^N E(I_X B_i' A_i B_i)$ and $S = \lim_{N\to\infty}\mathrm{Var}_{\mathbf{P}}(1/\sqrt{N}\sum_1^N I_X \times B_i' A_i \epsilon_i)$ with $B_i = (\partial/\partial\theta)h_0(X_i,\theta_0) + D_\theta r(h(X_i,\theta_0);\theta_0)$.

To prove Theorem 1, we show that the semiparametric WLS under Assumptions 1–8 satisfies Assumption A of Andrews (1994a).

**Assumption A(a).** This condition requires $\hat{\theta}$ to be a consistent estimator of $\theta_0$. Lemma 6 of Appendix A shows that under Assumptions 1–8, the semiparametric WLS is consistent.

**Assumption A(b).** The first part of condition A(b) is satisfied by Assumptions 3–7. Before we can show the second part of condition A(b) we need to specify a pseudometric $\rho_\Upsilon$ on $\Upsilon$. This pseudometric has to be the same as the one used to verify condition A(e). In this paper, we use the following pseudometric on $\Upsilon$:

$$\rho_\Upsilon(\tau_1,\tau_2) = \sup_{N^*\geq 1;\, i\leq N^*}[E\|m_i(\theta_0,\hat{\tau}) - m_i(\theta_0,\tau_0)\|^2]^{1/2}. \tag{A.8}$$

To prove condition A(b) we need to show $\rho_\Upsilon(\tau_0,\hat{\tau}) \xrightarrow{p} 0$. Using Minkowski's inequality applied to random vectors, it can be shown that

$$\begin{aligned}\rho_\Upsilon(\hat{\tau},\tau_0) \leq \sup_{N^*\geq 1;\, i\leq N^*}\Bigg\{&\Bigg[E\Bigg\|I_{X_N}\left(\frac{\partial}{\partial\theta}h_0(\theta_0,X_i) + D_\theta'\hat{r}(h(X_i,\theta_0);\theta_0)\right)\\ &\times A_i(r(h(X_i,\theta_0);\theta_0) - \hat{r}(h(X_i,\theta_0);\theta_0))\Bigg\|^2\Bigg]^{1/2}\\ &+ [E\|I_{X_N}(D_\theta'\hat{r}(h(X_i,\theta_0);\theta_0) - D_\theta' r(h(X_i,\theta_0);\theta_0))\\ &\times A_i(Y_i - h_0(\theta_0,X_i) - r(h(X_i,\theta_0);\theta_0))|^2]^{1/2}\Bigg\}.\end{aligned}$$

By repeated applications of Holder's inequality, to ensure condition A(b) it is sufficient to show that as $N\to\infty$

(b1) $\sup_{N^*\geq 1;\, i\leq N^*}[E\|I_{X_N}(r(h(X_i,\theta_0);\theta_0) - \hat{r}(h(X_i,\theta_0);\theta_0))\|^4]^{1/4} \xrightarrow{p} 0$.
(b2) $\sup_{N^*\geq 1;\, i\leq N^*}[E\|I_{X_N}(D_\theta\hat{r}(h(X_i,\theta_0);\theta_0) - D_\theta r(h(X_i,\theta_0);\theta_0))\|^4]^{1/4} \xrightarrow{p} 0$.
(b3) $E\|D_\theta\hat{r}(h(X_i,\theta_0);\theta_0)\|^8 < \infty$, $E\|(\partial/\partial\theta)h_0(\theta_0,X_i)\|^8 < \infty$, and $E\|\epsilon_i\|^8 < \infty$.

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

Recall $D_\theta r(h(X_i,\theta);\theta) = r^{(1)}(h(X_i,\theta);\theta)(\partial/\partial\theta')h(X_i,\theta_0) + r^{(2)}(h(X_i,\theta);\theta)$. Using standard inequalities it can be shown that for any $p > 1$

$$[E\|I_{X_N}(D'_\theta \hat{r}(h(X_i,\theta_0);\theta_0) - D'_\theta r(h(X_i,\theta_0);\theta_0))\|^p]^{1/p}$$

$$\leq [E\|I_{X_N}(\hat{r}^{(1)}(h(X_i,\theta_0);\theta_0) - r^{(1)}(h(X_i,\theta_0);\theta_0))\|^p]^{1/p}\left[E\left\|\frac{\partial h(X_i,\theta)}{\partial\theta'}\right\|^{2p}\right]^{1/2p}$$

$$+ [E\|I_{X_N}(\hat{r}^{(2)}(h(X_i,\theta_0);\theta_0) - r^{(2)}(h(X_i,\theta_0);\theta_0))\|^p]^{1/p}. \quad \textbf{(A.9)}$$

By Assumption 3,

$$E\left\|\frac{\partial h(X_i,\theta)}{\partial\theta'}\right\|^8 < \infty.$$

Then we can establish condition (b2) by showing that

$$[E\|I_{X_N}(\hat{r}^{(z)}(h(X_i,\theta_0);\theta_0) - r^{(z)}(h(X_i,\theta_0);\theta_0))\|^8]^{1/8} \xrightarrow{p} 0 \quad \textbf{(A.10)}$$

for $z = 1,2$. Because uniform consistency implies $L^Q$ consistency, the results in Lemma 7 are sufficient to ensure that (b1) and (b2) are satisfied. By Assumption 6, for large enough $N$ there exists a bounded subset of $\mathbf{R}^{l_j}, V_j^*$, such that the set $\{h_j(\theta_0, X_i) : X_i \in X_N\}$ is included in $V_j^* \cap V_j$.

**Assumption A(c).** This is an asymptotic orthogonality condition between the estimators $\hat{\theta}$ and $\hat{\tau}$. It is needed to show that if we use $\hat{\tau}$ instead of $\tau_0$ it will not affect the asymptotic distribution of $\hat{\theta}$. For the semiparametric WLS

$$\sqrt{N}\bar{m}_N^*(\theta_0,\hat{\tau}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N EI_X\Bigg(\frac{\partial}{\partial\theta}h_0(\theta_0,X_i) + D'_\theta\hat{r}(h(X_i,\theta_0);\theta_0))A_i$$

$$\times(\hat{r}(h(X_i,\theta_0);\theta_0) - r(h(X_i,\theta_0);\theta_0)\Bigg).$$

Using the fact $E[(\partial/\partial\theta)h_0(\theta_0,X_i) + D_\theta r(h(X_i,\theta_0);\theta_0)|h(X_i,\theta_0) = v] = \underline{0}\ \forall v$, Assumption A(c) holds if

$$\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N EI_X(D'_\theta\hat{r}(h(X_i,\theta_0);\theta_0) - D'_\theta r(h(X_i,\theta_0);\theta_0))A_i(\hat{r}(h(X_i,\theta_0);\theta_0) - r(h(X_i,\theta_0);\theta_0))\right\| \xrightarrow{p} 0.$$

By repeated applications of Holders's inequality

$$\|\sqrt{N}\bar{m}_N^*(\theta_0,\hat{\tau})\| \leq \frac{1}{\sqrt{N}}\sum_{i=1}^N (E\|I_X(D'_\theta\hat{r}(h(X_i,\theta_0);\theta_0) - D'_\theta r(h(X_i,\theta_0);\theta_0))\|^2)^{1/2}$$

$$\times\|A_i\|(E\|I_X(\hat{r}(h(X_i,\theta_0);\theta_0) - r(h(X_i,\theta_0);\theta_0))\|^4)^{1/4}.$$

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

Thus, to prove Assumption A(c) it is sufficient to show that

$$\sup_{i\leq N} N^{1/4}(E\|I_X(\hat{r}(h(X_i,\theta_0);\theta_0) - r(h(X_i,\theta_0);\theta_0))\|^4)^{1/4} \xrightarrow{p} 0$$

$$\sup_{i\leq N} N^{1/4}(E\|I_X(D_\theta \hat{r}(h(X_i,\theta_0);\theta_0) - D_\theta r(h(X_i,\theta_0);\theta_0))\|^2)^{1/2} \xrightarrow{p} 0.$$

By using the results of Lemma 7, we can conclude that A(c) holds.

**Assumption A(d).** Using Assumption 1, we can establish condition (d) by applying the central limit theorem (CLT) from Gallant (1987, p. 519) to $\nu_N(\tau_0)$, (in our case $\bar{m}_N^*(\theta_0,\tau_0) = 0$).

**Assumption A(e).** To verify this condition, we use the results of Theorem 4 in Appendix A. Note that the metric given in Theorem 4 is the same as the one in (A.8) because stochastic equicontinuity of a vector empirical process like ours follows from the stochastic equicontinuity of each element. Let

$$\mathcal{M} = \{m(\cdot,\theta_0,\tau) : \tau \in \mathcal{T}\},$$

where

$$m(\cdot,\theta_0,\tau) = I_{X_N}\left(\frac{\partial}{\partial\theta} h_0(\theta_0,X) + D_\theta \tau(v;\theta_0)\right)' A_i(Y - h_0(\theta_0,X) - \tau(v;\theta_0)).$$

Recall that $D_\theta\tau(v;\theta_0) = \tau^{(1)}(v,\theta_0)(\partial/\partial\theta)h(\theta_0,X) + \tau^{(2)}(v,\theta_0)$. By Assumptions 4–7, $\{\tau(v;\theta_0) : \tau \in \mathcal{T}\}$, $\{\tau^{(1)}(v;\theta_0) : \tau \in \mathcal{T}\}$, and $\{\tau^{(2)}(v;\theta_0) : \tau \in \mathcal{T}\}$ are type V classes of functions and satisfy Ossiander's $L^p$ entropy condition with envelopes $\sup_\tau|\tau(v;\theta_0)| < \infty$, $\sup_\tau|\tau^{(1)}(v;\theta_0)| < \infty$, and $\sup_\tau|\tau^{(2)}(v;\theta_0)| < \infty$. The expressions $\{Y - h_0(\theta_0,X)\}$, $\{(\partial/\partial\theta)h_0(\theta_0,X)\}$, $\{(\partial/\partial\theta)h(\theta_0,X)\}$, and $\{I(X_N)\}$ also satisfy Ossiander's $L^p$ entropy condition with envelopes $\|Y - h_0(\theta_0,X)\|$, $\|(\partial/\partial\theta)h_0(\theta_0,X)\|$, $\|(\partial/\partial\theta)h(\theta_0,X)\|$, and 1. Finally, by Theorem 6 of Andrews (1994b) we can conclude that $\mathcal{M}$ satisfies Ossiander's $L^p$ entropy condition if $(E\|Y - h_0(\theta_0,X)\|^8)^{1/8} < \infty$, $(E\|(\partial/\partial\theta)h_0(\theta_0,X)\|^8)^{1/8} < \infty$, and $(E\|(\partial/\partial\theta)h(\theta_0,X)\|^8)^{1/8} < \infty$.

**Assumption A(f).** Assume $m(\theta,\tau)$ and $M(\theta,\tau)$ exist uniformly over $(\theta,\tau)$ in $\Theta \times \Upsilon$. To show that $\{m_i(\theta,\tau)\}$ and $\{(\partial/\partial\theta')m_i(\theta,\tau)\}$ satisfy uniform WLLN's over $\Theta \times \Upsilon$, it is sufficient to show that (see Andrews, 1992):

(i) $\Theta \times \Upsilon$ is compact.
(ii) $m_i(\theta,\tau)$ and $(\partial/\partial\theta)m_i(\theta,\tau)$ are continuous in $\theta$ and $\tau$, uniformly over $\Theta \times \Upsilon$.
(iii) $P(\|1/N\sum_1^N m_i(\theta,\tau) - m(\theta,\tau)\| > \epsilon) \xrightarrow{p} 1$ for any $\theta \in \Theta$, $\tau \in \Upsilon$, and $\epsilon > 0$.
(iv) $P(\|1/N\sum_1^N (\partial/\partial\theta)m_i(\theta,\tau) - M(\theta,\tau)\| > \epsilon) \xrightarrow{p} 1$ for any $\theta \in \Theta$, $\tau \in \Upsilon$, and $\epsilon > 0$.

The compactness of $\Theta \times \Upsilon$ follows from our definition of $\Upsilon$ and Assumptions 1 and 2. A continuous function on a compact metric space is uniformly continuous on that space. Thus $m_i(\theta,\tau)$ and $(\partial/\partial\theta)m_i(\theta,\tau)$ are continuous in $\theta$ and $\tau$, uniformly over $\Theta \times \Upsilon$. Using the Kolmogorov weak law of large numbers and Assumption 1, we conclude that (iii) and (iv) are satisfied. Finally, the second part of Assumption A(f) requires continuity of $M(\theta,\tau)$ and $M(\theta,\tau)$ with respect to some pseudometrics for which $(\hat{\theta},\hat{\tau}) \xrightarrow{p} (\theta_0,\tau_0)$. If the following pseudometrics are used:

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press

$$\rho((\hat{\theta},\hat{\tau}),(\theta_0,\tau_0)) = \lim_{N^*\to\infty} \frac{1}{N^*} \sum_{1}^{N^*} E\|m_i(\hat{\theta},\hat{\tau}) - m_i(\theta_0,\tau_0)\|,$$

$$\rho((\hat{\theta},\hat{\tau}),(\theta_0,\tau_0)) = \lim_{N^*\to\infty} \frac{1}{N^*} \sum_{1}^{N^*} E\left\| \frac{\partial}{\partial\theta} m_i(\hat{\theta},\hat{\tau}) - \frac{\partial}{\partial\theta} m_i(\theta_0,\tau_0) \right\|,$$

the results of Lemma 7 are sufficient to ensure that both pseudometrics converge to zero. ■

https://doi.org/10.1017/S0266466600164047 Published online by Cambridge University Press